Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Pangenome analysis of human and aquaculture Streptococcus agalactiae identifies integrative conjugative elements as major carriers of antimicrobial resistance asymmetry

Sayed Mashequl Bari

bari.aahm@sau.edu.bd

Sher-e-Bangla Agricultural University

Fouzia Mostafa

Noakhali Medical College

Sabrina Hossain

WorldFish

Md. Abdullah Al Mamun

Sylhet Agricultural University

Muhammad Manjurul Karim

University of Dhaka

Research Article

Keywords:

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11578512/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Streptococcus agalactiae (Group B *Streptococcus*) causes neonatal sepsis in humans and is an emerging pathogen of farmed tilapia, yet how mobile genetic elements differ between these hosts is uncharacterized. We analyzed 259 genomes (159 human, 100 fish) from 25 countries, integrating pangenome-wide association, resistance gene profiling and coordinate-level mobile element analysis. Integrative conjugative elements (ICEs) were far more prevalent in human than fish genomes (73.6% versus 12.0%; odds ratio 20.43, $P = 1.72\times10^{-23}$), and human genomes carried more acquired resistance genes (median 1 versus 0; $P = 7.3\times10^{-26}$). Resistance genes lay inside ICE regions in 57.9% of human but 4.0% of fish genomes ($P = 4.4\times10^{-21}$); 55% of all resistance gene copies, and 81% of *tet*(M) copies, were ICE-located, versus 5% of random positions, although *erm*(A), *spw*, *lsa*(E) and *lnu*(B) never were. In fish, ICE-located resistance occurred in four genomes, three carrying *tet*(M) on Tn916-like elements. No β-lactam resistance gene was detected. Pangenome-wide association identified 27 host-associated gene families, with the adhesin *lmb* and four mobile-element genes the most robust. The zoonotic lineage ST283 occurred in both hosts. ICEs are the major identified carriers of acquired resistance in *S. agalactiae* and are sharply asymmetric across hosts, highlighting lineages for One Health surveillance.

Introduction

Streptococcus agalactiae (group B *Streptococcus*, GBS) is the leading cause of invasive neonatal disease worldwide, responsible for an estimated 231,000 infant cases, 46,000 stillbirths, and 91,000 infant deaths annually, with the greatest burden concentrated in low- and middle-income countries of sub-Saharan Africa and South and Southeast Asia [1, 2, 3]. Although intrapartum antibiotic prophylaxis reduces early-onset neonatal disease by approximately 80% where implemented, global coverage remains inconsistent, and no licensed maternal vaccine exists [3, 4]. The persistence of GBS as a leading cause of neonatal mortality despite decades of surveillance underscores the need for a systems-level understanding of its evolution and cross-population dissemination.

The significance of *S. agalactiae* extends beyond human medicine into a One Health imperative, as the same species causes devastating streptococcosis in farmed Nile tilapia (*Oreochromis niloticus*), with outbreak mortality reaching 80% and global aquaculture losses of hundreds of millions of US dollars annually [5, 6]. Critically, the intensive and often poorly regulated antimicrobial use characteristic of tilapia farming imposes concentrated selective pressure at the human–animal interface [7]. That this interface is more than theoretical was demonstrated unambiguously in 2015, when an outbreak of invasive disease among healthy adults in Singapore was genomically linked to consumption of raw freshwater fish and to sequence type 283 (ST283), a lineage indistinguishable between human cases and diseased tilapia [8, 9]. ST283 now accounts for a substantial share of adult invasive GBS across Southeast Asia [9], and phylogenomic evidence of human-to-animal host jumps in this species [10] shows that GBS readily crosses host boundaries, raising the question of whether resistance crosses with it.

The capacity of *S. agalactiae* to colonize hosts as disparate as humans and fish is rooted in the plasticity of its accessory genome, sustained by one of the most dynamic mobilomes among Gram-positive pathogens. GBS holds a foundational place in bacterial genomics as the species in which the pangenome concept was first defined, with its total gene repertoire shown to be functionally open [11]. This openness is driven by a mobilome of integrative and conjugative elements (ICEs), prophages, insertion sequences, and plasmids that mediate horizontal gene transfer across lineages, hosts, and geographies [12]. ICEs are major contributors to GBS genome diversity and the major vehicles of large-scale horizontal transfer in the species, integrating site-specifically into conserved chromosomal loci and mobilizing resistance and adaptation determinants across species and genus boundaries [13]. The evolutionary weight of this machinery is established: Da Cunha *et al.* showed that acquisition of a Tn916-family ICE carrying *tet*(M) drove a rapid clonal sweep in the 1960s that replaced the ancestral GBS population with the resistance-bearing human-adapted

lineages that dominate clinical disease today [14]. Whether comparable ICE-driven dynamics now operate at the human–aquaculture interface is the central question this study addresses.

Recent large-scale comparative analyses have advanced understanding of GBS host adaptation while defining the limits of current knowledge. Gori et al. performed the most comprehensive pan-GWAS of GBS to date, resolving 279 clonal-complex-associated accessory gene families across nearly 2,000 genomes [15]; however, this analysis was restricted to human isolates and therefore cannot address the genomic determinants of cross-host adaptation or the distribution of mobile resistance elements between host compartments. Crestani et al. extended comparative genomics to a multi-host framework of 1,254 human, bovine, and fish genomes, defining three host-associated gene clusters and establishing that host-generalist and host-restricted sub lineages coexist within the species [16] a finding that directly frames ST283 as a candidate generalist lineage. Yet neither study performed coordinate-level mobilome characterization, systematically assessed AMR gene co-localization with ICE boundaries, or applied population-structure-corrected pan-GWAS across host compartments. Fish-adapted ST260–261 lineages, by contrast, have undergone marked reductive genome evolution during piscine specialization [17], consistent with the principle that host restriction accompanies accessory-genome contraction while host generalism requires its expansion. Bovine GBS, a third major compartment with distinct gene content [18], falls outside the present scope; we focus on the human–aquaculture interface, where the ST283 linkage, overlapping antimicrobial pressures, and shared geography provide the most tractable context for investigating ICE-mediated resistance exchange.

To address these questions, we conducted a comparative One Health pangenome analysis of 259 *S. agalactiae* genomes from human clinical and fish (*O. niloticus*) sources, spanning 25 countries across five continents and subjected to rigorous quality control for completeness, contamination, and chimerism. To our knowledge, this is the most comprehensive dual-host GBS pangenome investigation to date, and the first to integrate coordinate-level mobilome characterization with population-structure-corrected host-association analysis. Our objectives were threefold: to characterize and compare the pangenome architectures, phylogenomic structure, and host-associated gene content of human- and fish-adapted GBS populations; to profile their antimicrobial resistance and virulence repertoires and quantify host association while controlling for false discovery; and to characterize the full mobilome ICEs, prophages, insertion sequences, and plasmids resolving the physical co-localization of resistance determinants with ICE boundaries and identifying ICE families shared across the human–aquaculture interface. Together, these analyses delineate the genomic architecture of host adaptation and the mobilome's role in the distribution of resistance distribution across host compartments, providing an evidence base for integrated One Health surveillance of GBS.

Methods

Genome dataset assembly and host metadata curation

All *S. agalactiae* genome assemblies deposited in the NCBI Assembly database (taxonomy identifier 1311) were retrieved on 31 August 2026 using the NCBI Datasets command-line interface (v18.36.0) [19] and the Biopython Entrez interface (v1.84) [20]. The initial query targeted Complete Genome, Chromosome, and Scaffold assembly levels; Contig-level assemblies were not explicitly queried, but a subset (n = 18; 10 human, 8 fish) retrieved under the Scaffold query were reclassified as Contig-level in the NCBI Assembly database at the time of download. All 18 passed quality thresholds and were retained, with contig proportions balanced between host groups (human 6.3%, fish 8.0%; mean N50: human 73,305 bp, fish 76,971 bp). Host and isolation-source metadata were extracted from linked NCBI BioSample records via Entrez, and each genome was classified as human-associated, fish-associated, or other based on host field annotation, with ambiguous records resolved by manual inspection. Fish-associated genomes were retained without subsampling to maximize representation of piscine lineages. Human-associated genomes underwent stratified country-of-origin subsampling via round-robin country rotation (random seed = 42, target n = 180) to maximize geographic

diversity while preventing overrepresentation of any single country. This yielded an initial dataset of 286 candidate genomes (180 human, 106 fish) from 25 countries across five continents.

Taxonomic confirmation and assembly quality control

Taxonomic identity of all 286 assemblies was confirmed as *S. agalactiae* using GTDB-Tk v2.7.2 (classify_wf workflow) against the Genome Taxonomy Database release R232, with skani v0.3.2 for ANI-based pre-screening [21]. Assembly completeness and contamination were assessed with CheckM2 v1.1.0 (Apptainer container; reference database: Zenodo DOI 10.5281/zenodo.14897628) [22]. Chimeric assemblies were identified with GUNC v1.0.6 against the progenomes2.1 database [23]. Assembly statistics were computed with SeqKit v2.13.0 [24]. Genomes meeting all three criteria (CheckM2 completeness ≥ 98%, contamination ≤ 3%, and GUNC pass) were retained. Of the 286 candidates, 259 passed (159 human, 100 fish), while 27 were excluded (26 for CheckM2 failure; one, GCF_001074585.1, for chimera detection by GUNC, clade separation score 0.79). Retained genomes exhibited a mean completeness of 99.96% (range 98.38–100.00%) and a mean contamination of 0.305% (range 0.00–2.81%). Dereplication at 99.9% ANI was evaluated using dRep v3.7.1 [25]. However, ST7 isolates from five Latin American countries showed ≥ 99.9% ANI, indicating genuine geographic diversity rather than technical redundancy. Consequently, all 259 post-QC genomes were retained without dereplication.

Genome annotation, sequence typing, and serotyping

All 259 genomes underwent uniform re-annotation using Bakta v1.12.1 (database v6.0, released 2025-02-24; Zenodo DOI 10.5281/zenodo.14916843) [26] to eliminate heterogeneity arising from disparate NCBI annotation pipelines, which would otherwise introduce systematic gene-calling inconsistencies into pangenome reconstruction. The integrated AMRFinderPlus resistance gene database within Bakta was updated to v2026-08-07.1 immediately prior to annotation using AMRFinderPlus, a tool that identifies AMR genes, resistance-associated point mutations, developed by NCBI. Bakta GFF3 output files served as input for all downstream pangenome analyses; Bakta GBFF output was used for prophage prediction. Multilocus sequence types were assigned against the *S. agalactiae* PubMLST scheme using mlst v2.35.0 [27, 28], and clonal complexes were inferred from established allelic relationships in the PubMLST database. Capsular serotypes were determined in silico using GBS-SBG (github.com/swainechen/GBS-SBG) [29], a GBS-specific serotyping tool employing BLAST v2.17.0 + to screen assembled genomes against capsular biosynthesis locus sequences; the highest-scoring serotype assignment was reported per genome. Genomes yielding no unambiguous serotype assignment were recorded as non-typeable.

Pangenome reconstruction and functional analysis

The pangenome was reconstructed from Bakta GFF3 annotation files for all 259 genomes using Panaroo v1.8.0 [30], which applies graph-based correction of annotation errors and fragmented gene calls prior to gene family clustering. Gene families were partitioned into four classes: core (present in ≥ 99% of genomes), soft-core (95–99%), shell (15–95%), and cloud (< 15%). The strict core threshold of 99% was chosen to minimize the influence of annotation artifacts on core genome definition; this threshold yields a smaller but higher-confidence core than the more commonly used 95% cutoff and should be considered when comparing core genome sizes across studies. Pangenome openness was assessed by fitting Heaps' law to gene-accumulation curves generated from 100 random genome-addition permutations; the decay parameter α was estimated by log-linear regression of mean cumulative gene family count against genome number. A value of $\alpha < 1$ indicates an open pangenome. Pangenome summary statistics including total pangenome size, core genome size, and partition proportions were extracted from Panaroo output and benchmarked against prior GBS pangenome analyses where comparable parameters were reported.

COG functional categories were taken from the COG category letters in the Bakta cross-reference field ('COG:<letter>') of every coding sequence in the 259 genomes; a gene assigned to several categories was counted once in each. For host comparison, the number of genes in each category was expressed as a proportion of all coding sequences of the genome (to remove the effect of genome size) and compared between hosts with a two-sided Mann–Whitney U test per category, with Benjamini–Hochberg correction over all categories containing at least one gene. KEGG Orthology (KO) identifiers were extracted from Bakta cross-reference fields, yielding 564 unique KO terms across the dataset; genome-level prevalence (proportion of genomes carrying ≥ 1 copy) was calculated per host for each KO, and the most host-differential terms by prevalence difference (human minus fish) were visualized. All pangenome statistics and accumulation curves were visualized in R v4.5.3 using ggplot2 v3.5.0 and ComplexHeatmap v2.20.0 [31].

Core-genome phylogeny and average nucleotide identity analysis

Recombinant regions in the Panaroo core-genome alignment were identified and masked using Gubbins v3.4.1 [32]. The recombination-filtered SNP alignment comprised 259 sequences across 34,464 alignment columns (27,132 parsimony-informative; 7,332 singleton sites); 1,156,966 constant sites were reconstructed from the full core alignment (proportion of invariable sites = 0.972) and incorporated into likelihood calculations using the -fconst correction. Maximum-likelihood phylogeny was inferred using IQ-TREE v3.1.3 [33] under the GTR + G4 + I substitution model selected by ModelFinder Plus using the Bayesian Information Criterion; branch support was estimated from 1,000 ultrafast bootstrap replicates with NNI optimization, with nodes achieving ultrafast bootstrap ≥ 95% considered robustly supported. The final tree achieved LogL = − 434,869.899; of 248 internal branches, 142 (57%) had ultrafast bootstrap support ≥ 95%, including the branches that define the ST7, ST283, ST261 and ST927 lineages and the two fish-only clades (all 100%). Phylogenetic annotations encompassing host category, clonal complex, sequence type, capsular serotype, continent of isolation, and presence of key resistance determinants (tet(M), erm(B)) and mobile genetic elements (ICE, prophage count) were prepared as iTOL-formatted annotation files and visualized using the Interactive Tree of Life (iTOL) v6 [35]. The recombination-filtered phylogeny additionally served as the basis for constructing the Bray-Curtis distance-based kinship matrix used in pan-GWAS linear mixed model analysis to control for clonal population structure. Lineage monophyly was assessed on the unrooted tree (a group was treated as a clade if a single branch separated it from all other genomes), and host structure was quantified by Fitch parsimony, comparing the minimum number of host changes on the tree with that obtained from 1,000 random permutations of the host labels. Pairwise average nucleotide identity was calculated between all genomes using FastANI v1.34 [34] (66,822 directional comparisons, 33,411 genome pairs); all assemblies exceeded the 95% species-boundary threshold confirming species-level identity. ANI values were summarized for three comparison groups (within-human, within-fish, cross-host); because pairwise values are not independent, differences between groups are reported descriptively without significance tests. The statistical significance of accessory genome size differences between host groups was assessed by Wilcoxon rank-sum test.

Antimicrobial resistance and virulence gene profiling

Acquired antimicrobial resistance genes were identified across all 259 genomes using AMRFinderPlus v4.2.7 (organism *Streptococcus*; database v2026-08-07.1) [36], which screens against a curated hierarchy of resistance gene families, point mutations, and stress response elements. Virulence-associated genes were detected using ABRicate v1.4.0 against the Virulence Factor Database (VFDB; core dataset, downloaded 3 September 2026) [37], applying minimum thresholds of ≥ 80% nucleotide identity and ≥ 80% query coverage, consistent with VFDB recommended screening parameters for conserved virulence factor detection across divergent streptococcal lineages. For each genome, AMR and virulence gene presence/absence were recorded as a binary matrix, and total gene counts per genome were computed for burden analysis; genes detected in fewer than two genomes across the entire dataset were excluded from

prevalence-based comparative analysis to remove sequencing or annotation artifacts while retaining rare but potentially epidemiologically significant determinants. Between-host differences in gene prevalence were assessed by Fisher's exact test, with Benjamini–Hochberg correction applied across all genes tested simultaneously to control the false discovery rate; genes with adjusted $P < 0.05$ were considered significantly host-associated. Between-host differences in total gene burden per genome were assessed by Mann–Whitney U test. Odds ratios with 95% confidence intervals were computed using the scipy.stats and statsmodels libraries in Python v3.10. Results were visualized as hierarchically clustered presence/absence heatmaps (Ward's D2 linkage) using ComplexHeatmap v2.20.0 [31] in R v4.5.3, with genomes split by host group and gene rows annotated by antibiotic class (AMR) or functional category (virulence). Prevalence bar charts per host group were appended as row annotations to facilitate quantitative comparison alongside the binary heatmap.

Pangenome-wide association analysis

Host-associated gene families were identified using Pyseer v1.4.0 with a linear mixed model [38]. The Panaroo gene presence/absence matrix was used as the genotype input, and host category was encoded as a binary phenotype (human = 1, fish = 0). Population structure was inferred from a kinship matrix derived from pairwise phylogenetic distances of the recombination-filtered IQ-TREE maximum-likelihood tree, normalized to the maximum observed distance. Of 6,136 gene families loaded, 2,811 were tested after pre-filtering, with a Bonferroni-corrected significance threshold of $P < 1.78\times10^{-5}$ ($\alpha = 0.05/2{,}811$). The estimated genomic heritability ($h^2=1.00$) indicates that host category is near-completely explained by the kinship matrix, reflecting the tight coupling of host and clonal background in this dataset (ST7 and ST260 are fish-exclusive; CC1, CC17, and CC19 are human-enriched), and should be interpreted as confirmation that the LMM adequately controls for population structure rather than as evidence of biological heritability in the classical sense. Gene families that failing the Pyseer chi-squared goodness-of-fit filter (bad-chisq), which identifies results where allele frequency distributions deviate significantly from model assumptions, were flagged and excluded from the final set of significant associations. Significant gene families were further classified by concordance between allele-frequency direction and effect-size (β) sign: families with AF and β pointing in the same direction were classified as robust associations, while discordant families were interpreted as lineage markers reflecting host–clade confounding. Significant gene families were annotated using Bakta TSV output and assigned to functional categories based on predicted protein function.

Mobilome characterization and resistance-element co-localization

IS elements were catalogued across all 259 genomes using ISEScan v1.7.3 [39]. ICEs were identified by BLASTn screening (BLAST+ v2.17.0; E-value ≤ 10^{-10}; ≥80% nucleotide identity) against the ICEberg 2.0 database (554 curated ICE sequences) [40]; hits with alignment length ≥ 15 kb and nucleotide identity ≥ 90% were classified as confirmed ICEs. ICE regions are expressed in the coordinates of the genome assembly. Because multiple closely related ICEberg reference sequences frequently matched the same genomic region, overlapping ICE coordinate hits within each genome were merged into single non-redundant ICE regions prior to downstream analysis. Prophage regions were predicted with PhiSpy v5.0.10 using Bakta GBFF annotation as input [41]. Plasmid replicons and conjugative mobility were characterized with MOB-suite v3.1.9 (mob_recon) [42]. AMR gene co-localization with ICEs was assessed at coordinate resolution: an acquired resistance gene was classified as ICE-associated only when its genomic coordinates fell entirely within the boundaries of a merged ICE region in the same genome. To estimate the proportion expected by chance, 20 random positions were drawn on the contig of every resistance gene and tested in the same way. The ICEberg references matched by each region were used to describe its type. Cross-host ICE sharing was defined as detection of an identical ICE family, identified by its ICEberg database identifier, in genomes from both host groups. Between-host differences in MGE prevalence were assessed by Fisher's exact test.

Results

Dataset composition and host-associated population structure

A total of 286 *Streptococcus agalactiae* genome assemblies were retrieved from the NCBI Assembly database (31 August 2026) using the NCBI Datasets CLI v18.36.0, comprising 180 human-associated and 106 fish-associated isolates (predominantly *Oreochromis niloticus*). All 286 genomes were confirmed as *S. agalactiae* by GTDB-Tk v2.7.2. Following quality control (CheckM2 completeness ≥98%, contamination ≤3%, GUNC chimera screening), 259 genomes were retained (159 human, 100 fish) and 27 excluded (26 CheckM2 failure, 1 GUNC-detected chimera; Table S1). Retained genomes had a mean completeness of 99.96% and a mean contamination of 0.31% (Table S1).

Of the 259 retained genomes, 215 carried country-of-origin metadata, spanning 25 countries across five continents (North America, South America, Asia, Europe, and Africa); geographic origin was not recorded for the remaining 44 genomes (Fig. 1a). Human-associated genomes originated from 15 countries across all five continents, whereas fish-associated genomes originated from 14 countries restricted to three continents (Asia, North America, and South America), with no fish-associated genomes available from Africa or Europe (Table 1). Human-associated genomes were significantly larger than fish-associated genomes (mean 2.14 Mb versus 1.99 Mb; Wilcoxon rank-sum test, $P=1.5\times10^{-12}$; Fig. 1d).

Multilocus sequence typing (mlst v2.35.0, PubMLST scheme) resolved 222 of 259 genomes to distinct sequence types; 37 genomes, predominantly fish-associated draft assemblies, yielded incomplete allele assignments due to locus fragmentation. The two host populations were sharply differentiated in clonal composition (Fig. 1b; Table 2). Fish-associated genomes were dominated by ST7 (n=51) and ST260 (n=14), both fish-exclusive. Human-associated genomes were distributed across ST1 (n=29), ST12 (n=16), ST19 (n=12), ST23 (n=9), ST459 (n=8), and ST17 (n=8; CC17), spanning clonal complexes CC1, CC17, CC19, and CC23. ST283 was the only sequence type with substantial representation in both host groups (human n=8, fish n=2).

Capsular serotyping (GBS-SBG) resolved 11 serotypes across the dataset, with three human genomes non-typeable (Fig. 1c; Table 2). Fish-associated genomes were dominated by serotypes Ia (n=51) and Ib (n=46), together comprising 97% of piscine genomes. Human-associated genomes were more serotypically diverse, with the most common serotypes being V (n=40), II (n=24), III-1 (n=13), and III-2 (n=9); III-2 was detected exclusively in human genomes. Serotype III-4, tracking the cross-host ST283 lineage, was present in both groups (human n=8, fish n=2).

Table 1. Summary of the *S. agalactiae* genome dataset analyzed in this study, stratified by host of isolation.

Host category	Genomes (n)	Countries (continents)	Mean genome size (Mb)	Predominant ST	Predominant serotype
Human	159	15 (5)	2.14	ST1, ST12, ST17, ST19, ST23	V, II
Fish	100	14 (3)	1.99	ST7, ST260	Ia, Ib

Table 2. Sequence types (STs), clonal complexes (CCs) and capsular serotypes of the 259 genomes by host. STs with at least three genomes are listed individually; NT, non-typeable; —, none assigned.

Sequence type	Clonal complex	Human genomes	Fish genomes	Capsular serotypes (n)
ST7	CC7	0	51	Ia (51)
ST1	CC1	29	1	V (21), Ib (5), VI (3), VIII (1)
ST12	CC1	16	0	Ib (14), NT (1), II (1)
ST260	CC260	0	14	Ib (14)
ST19	CC19	12	0	III-1 (10), V (2)
ST283	CC283	8	2	III-4 (10)
ST23	CC23	9	0	Ia (7), III-3 (2)
ST17	CC17	8	0	III-2 (7), NT (1)
ST459	CC459	8	0	IV (8)
ST103	CC103	5	0	Ia (4), III-3 (1)
ST452	—	5	0	IV (5)
ST591	—	5	0	II (5)
ST929	—	4	0	V (4)
ST2	CC1	3	0	V (3)
ST10	CC10	3	0	Ib (2), II (1)
ST22	CC22	3	0	II (3)
ST88	—	3	0	II (3)
ST832	—	3	0	V (3)
ST927	—	0	3	Ib (3)
Other STs (22 STs, 1–2 genomes each)	—	25	2	—
ST not resolved	—	10	27	—
Total		159	100	

Pangenome architecture and host-differential gene content

Pangenome reconstruction across 259 *S. agalactiae* genomes identified 6,136 gene families, partitioned into 1,524 core (≥99% of genomes; 24.8%), 56 soft-core (95–99%; 0.9%), 712 shell (15–95%; 11.6%) and 3,844 cloud (<15%; 62.7%) families (Fig. 2a). The gene-frequency distribution was strongly U-shaped, and Heaps' law modelling of gene-accumulation curves yielded a decay parameter $\alpha = 0.211$, confirming an open pangenome in which the pan-genome continued to expand while the core genome stabilized at 1,524 families (Fig. 2b, 2c). Human-associated genomes carried a significantly larger accessory genome than fish-associated genomes (mean 527 versus 377 gene families per genome; Mann–Whitney U test, $U = 12{,}570$, $P = 3.5 \times 10^{-15}$).

COG categories were assigned to 72.2% of coding sequences in human and 70.9% in fish genomes (mean 1,503 and 1,393 per genome; Table S2). Functional composition was nearly identical between hosts: across the 21 informative categories, the share of coding sequences differed by at most 0.35 percentage points (carbohydrate transport and metabolism, 7.19% versus 6.84%), and $\log_2$ ratios ranged from −0.09 to +0.09 (Fig. 2d). Thirteen categories differed significantly after Benjamini–Hochberg correction, 11 slightly higher in human genomes (for example cell-cycle control, carbohydrate metabolism and defense mechanisms, $\log_2$ ratio +0.07 each) and two slightly higher in fish genomes (cell motility and coenzyme metabolism). Human genomes carried more genes in every informative category except cell motility, consistent with their larger gene content (mean 2,082 versus 1,964 coding sequences per genome). Categories A and B, which contain about one gene per genome, were not interpreted.

KEGG Orthology analysis identified 564 unique KO terms (Fig. 2e; Table S3). The most host-differential terms revealed contrasting mobile-element signatures between compartments: the insertion-sequence IS3-family transposase K07487 (present in 43.0% of fish versus 6.3% of human genomes) and a further transposase K07494 (42.0% of fish versus 0% of human genomes) were markedly enriched in fish-associated genomes, whereas the human-restricted KO K18220 was present in 73.0% of human versus 3.0% of fish genomes.

Phylogenomic relationships and host-associated genomic divergence

The recombination-filtered core-genome phylogeny showed strong host structure (Fig. 3a; Table S4): parsimony reconstruction required four host changes on the tree, compared with 69–91 when host labels were permuted ($P \leq 0.001$). Fish genomes fell into two fish-only clades an ST7 clade (n=51; all serotype Ia) and a clade of 46 genomes containing ST260, ST927, ST261 and genomes with unresolved ST (all serotype Ib), each supported at 100% which together held 97 of the 100 fish genomes; the remaining three (two ST283 genomes and one ST1 genome) sat within lineages that also contain human genomes. Human-associated genomes were distributed across the remaining lineages, including CC1, CC17, CC19 and CC23, and all eight ST17 genomes (CC17) fell within a single exclusively human clade of ten genomes (100%; two genomes had unresolved ST). ST283 (human n=8, fish n=2) was detected in both host compartments and formed one clade of ten genomes (100%) that included isolates of both hosts. Support was not uniformly high: 142 of 248 internal branches (57%) had ultrafast bootstrap support ≥95%, and the branch linking the Bangladeshi fish ST1 genome to its nearest human relatives had 77%.

Pairwise average nucleotide identity across all 66,822 directional genome comparisons (33,411 genome pairs) confirmed that all 259 assemblies fell within a single species-level cluster, with ANI ranging from 98.00% to 100.00%, above the 95% species boundary (Fig. 3b). Mean ANI was higher within fish genomes (99.23%; n=9,900 comparisons) and within human genomes (99.05%; n=25,122) than between hosts (98.88%; n=31,800), a small (0.17–0.35 percentage-point) but consistent shift indicating modest host-associated genomic divergence below the species boundary.

Human-associated genomes carried significantly more accessory gene families than fish-associated genomes (mean 527 versus 377 per genome; median 505 versus 423; Mann–Whitney U test, $U = 12{,}570$, $P = 3.5 \times 10^{-15}$; Fig. 3c).

Antimicrobial resistance and virulence gene repertoires

The total burden of acquired resistance and virulence genes per genome varied markedly among host. Human genomes carried a median of 1 acquired resistance gene (mean 1.93, range 0–10) compared to a median of 0 in fish genomes (mean 0.10, range 0–4; Mann–Whitney U test, $P = 7.3 \times 10^{-26}$; Fig. 4a), with 131 of 259 genomes (94 fish, 37 human) carrying no acquired resistance gene at all. Total virulence gene burden was similarly higher in human genomes (median 37, range 13–44) than fish genomes (median 28, range 14–39; $P = 7.0 \times 10^{-22}$; Fig. 4b).

Resistance gene carriage was markedly asymmetric between host groups: 12 of 22 detected gene families were significantly human-enriched after Benjamini–Hochberg correction (adj. $P < 0.05$; Fig. 4c; Table 3). The tetracycline

determinant tet(M) was the most prevalent and most differentially distributed gene, detected in 91/159 human genomes (57.2%) versus 3/100 fish (3.0%) genomes (OR = 43.3; adj. P = 1.43×10^{-20}). The macrolide–lincosamide–streptogramin B determinant erm(B) was also human-enriched (40/159, 25.2% versus 2/100, 2.0%; OR = 16.5; adj. P = 1.58×10^{-6}). Seven further resistance genes tet(O), mef(A), msr(D), aadE, spw, lsa(E) and lnu(B) were detected exclusively in human genomes (adj. P < 0.05), and the aminoglycoside determinants *ant*(6)-Ia and *aph*(3')-IIIa and the macrolide–lincosamide determinant *erm*(A) were each present in 15–22 human genomes but ≤1 fish genome. No acquired β-lactam resistance gene was detected in any genome of either host group. The complete unfiltered AMRFinderPlus output is provided in Table S5.

Virulence gene screening identified 8,146 hits across 45 gene families with three distinct distribution patterns (Fig. 4c; Table S6). A core set was near-universal in both hosts: capsule biosynthesis genes (neuA, neuB, neuC; 100% in both), CAMP factor (cfb; 100% in both) and hyaluronidase (hylB; 89.3% human, 100% fish). A second set was strongly human-enriched: the laminin-binding protein gene lmb (133/159, 83.6% versus 3/100, 3.0%; Fisher P = 2.4×10^{-42}), pilus-island genes pilC, srtC3 and srtC4 (75.5% versus 3.0%), the C5a peptidase gene *scpA/scpB* (67.9% versus 2.0%), and the cytolysin operon (cylD; 97.5% versus 54.0%). A single gene, GBS_RS06570, showed fish-specific enrichment (99.0% fish versus 47.8% human; Fisher P = 1.2×10^{-21}).

Table 3. Prevalence and host association of acquired antimicrobial resistance genes (n = 259). Antibiotic class as assigned by AMRFinderPlus; OR, odds ratio (>100 where one cell is zero); adj. P, Benjamini–Hochberg-adjusted Fisher's exact P.

AMR gene	Antibiotic class	Human, n (%)	Fish, n (%)	OR	adj. P (BH)
tet(M)	Tetracycline	91 (57.2)	3 (3.0)	43.3	1.43×10^{-20}
erm(B)	Lincosamide/macrolide/streptogramin	40 (25.2)	2 (2.0)	16.5	1.58×10^{-6}
tet(O)	Tetracycline	24 (15.1)	0 (0.0)	>100	5.40×10^{-5}
ant(6)-Ia	Aminoglycoside	22 (13.8)	1 (1.0)	15.9	6.76×10^{-4}
mef(A)	Macrolide (efflux)	21 (13.2)	0 (0.0)	>100	1.42×10^{-4}
msr(D)	Macrolide/streptogramin	20 (12.6)	0 (0.0)	>100	2.14×10^{-4}
aadE	Aminoglycoside	15 (9.4)	0 (0.0)	>100	1.97×10^{-3}
erm(A)	Lincosamide/macrolide	15 (9.4)	1 (1.0)	10.3	1.53×10^{-2}
aph(3')-IIIa	Aminoglycoside	15 (9.4)	1 (1.0)	10.3	1.53×10^{-2}
spw	Aminoglycoside	10 (6.3)	0 (0.0)	>100	1.72×10^{-2}
lsa(E)	Lincosamide/streptogramin	9 (5.7)	0 (0.0)	>100	2.77×10^{-2}
lnu(B)	Lincosamide	8 (5.0)	0 (0.0)	>100	4.56×10^{-2}
β-lactam genes	β-lactam	0 (0.0)	0 (0.0)	—	—

Pangenome-wide association analysis identifies host-specific gene families

Pangenome-wide association analysis was performed using a linear mixed model to identify gene families associated with host of isolation while controlling for clonal population structure. Host category and clonal background were strongly confounded in this dataset, with fish-associated genomes largely restricted to ST7 and ST260; consistent with this confounding, 2,285 of 2,811 tested gene families (81%) failed the model's chi-squared goodness-of-fit filter (bad-chisq). The associations reported here are therefore interpreted as gene families that distinguish the two host

compartments rather than as experimentally validated determinants of host-adaptation. After excluding bad-chisq-flagged families and applying Bonferroni correction ($P < 1.78 \times 10^{-5}$), 28 gene families reached significance, all more prevalent in human than fish genomes (Fig. 5; Table 4). One family, group_2481, showed near-identical allele frequencies between hosts (0.054 versus 0.053) with an effect size discordant with its allele frequency and was flagged as a residual population-structure artifact, leaving 27 host-associated families: 17 prophage-associated, four mobile-element, four of unknown function, one adhesin, and one transporter. The full unfiltered results are provided in Table S7.

Six of these families showed concordant allele-frequency and effect-size direction, representing the most robust host associations. Notably, four were mobile-element genes two IS3-family transposases (group_3148, $P = 7.67 \times 10^{-9}$; group_3012, $P = 2.16 \times 10^{-7}$), an IS150-family element (group_4309, $P = 2.93 \times 10^{-12}$), and a UPF0236 transposase-like family (group_3273, $P = 9.58 \times 10^{-14}$) each exclusively detected in human genomes. The two remaining robust associations were the laminin-binding adhesin lmb ($P = 9.32 \times 10^{-16}$; allele frequency 0.523 in human versus 0.000 in fish), independently corroborated by VFDB screening (83.6% versus 3.0%; Fig. 4c), and an S8-family serine peptidase (group_3283). The remaining 21 significant families showed effect-size directions discordant with their allele frequencies, consistent with confounding between host and clonal background, and are more parsimoniously interpreted as lineage markers. Among these, 13 prophage structural gene families encoding capsid, portal, terminase, tail and morphogenesis components shared identical allele frequencies (human 0.089, fish 0.002) and P values (4.53×10^{-9}), indicating a single co-inherited human-associated prophage rather than independent signals; this mirrors the prophage prevalence asymmetry independently detected by PhiSpy (100% in human versus 54.0% in fish genomes).

Table 4. Gene families significantly associated with host of isolation (2,811 gene families tested; Bonferroni threshold $P < 1.78 \times 10^{-5}$). P values are likelihood-ratio test P values. Reliability: robust, allele-frequency and effect-size directions agree; lineage-confounded, they disagree; artifact, near-identical allele frequencies in the two hosts. *Cluster znuA~~~lmb. †Thirteen co-inherited families with identical allele frequencies and P values (group_3917, group_3916, group_3864, group_3847, group_3833, group_3585, group_3545, group_3544, group_3445, group_3408, group_3373, group_3298 and group_2190).

Gene / cluster	Predicted function (Bakta)	Category	AF human	AF fish	β	P (LRT)	Reliability
group_2481	Phage protein	Prophage	0.054	0.053	−0.615	6.73×10^{-26}	Artifact
*lmb**	Laminin-binding surface protein (Lmb)	Adhesion	0.523	0.000	+0.385	9.32×10^{-16}	Robust
group_3273	UPF0236 transposase-like protein	Mobile element	0.519	0.000	+0.336	9.58×10^{-14}	Robust
group_3283	S8 family serine peptidase	Unknown	0.519	0.000	+0.336	9.58×10^{-14}	Robust
group_4309	IS150 InsJ-like HTH protein	Mobile element	0.671	0.000	+0.300	2.93×10^{-12}	Robust
group_2160	Hypothetical phage protein	Prophage	0.070	0.012	−0.302	3.44×10^{-12}	Lineage-confounded
group_2988	Winged HTH domain protein	Unknown	0.814	0.000	−0.445	5.87×10^{-10}	Lineage-confounded
ykoI	PepSY domain-containing protein	Unknown	0.814	0.000	−0.445	5.87×10^{-10}	Lineage-confounded
group_3533	HicB-like antitoxin	Unknown	0.085	0.003	−0.224	3.24×10^{-9}	Lineage-confounded
group_3931	Terminase small subunit	Prophage	0.085	0.003	−0.223	3.78×10^{-9}	Lineage-confounded
group_3614	PBSX terminase large subunit	Prophage	0.085	0.003	−0.223	3.78×10^{-9}	Lineage-confounded
group_3917–group_3298†	Prophage structural cluster (13 families)	Prophage	0.089	0.002	−0.221	4.53×10^{-9}	Lineage-confounded
group_4268	MerR-type HTH protein	Prophage	0.333	0.000	−0.310	6.49×10^{-9}	Lineage-confounded
group_3148	IS3 family transposase ORF B	Mobile element	0.473	0.000	+0.211	7.67×10^{-9}	**Robust**
group_2671	ABC transporter ATP-binding protein	Transport	0.752	0.000	−0.261	6.13×10^{-8}	Lineage-confounded
group_3012	IS3 ISSag2 transposase ORF A	Mobile element	0.434	0.000	+0.173	2.16×10^{-7}	**Robust**

Mobilome architecture and its association with antimicrobial resistance

The mobilome of the 259 *S. agalactiae* genomes was characterized across four categories of mobile genetic element (Fig. 6). Plasmids were rare and non-differentially distributed between host groups (human 8/159, 5.0%; fish 7/100, 7.0%; Fisher's exact test, P = 0.61). In contrast, prophage burden was strongly host-asymmetric: prophages were detected in all 159 human genomes (100%; mean 3.12 per genome) versus 54/100 fish genomes (54.0%; mean 1.04 per genome; $P = 2.77 \times 10^{-23}$; Fig. 6a). Insertion-sequence profiling identified IS3 as the dominant family in both hosts (human 140 copies, fish 56 copies), with host-partitioned minor families: ISL3, IS256 and ISAS1 were detected only in human genomes, and IS982 and IS630 only in fish genomes (Fig. 6b).

Integrative conjugative elements showed the largest host-differential prevalence of any mobile element category, detected in 117/159 human (73.6%) versus 12/100 fish (12.0%) genomes (OR = 20.43, P = 1.72×10^{-23}; Fig. 6a). Of 65 ICEberg reference entries matched, 34 were detected in genomes from both host groups and 31 in human genomes only; no entry was exclusive to fish (Fig. 6c; Table S8). Because closely related references match the same regions, these entries are not independent elements; merging overlapping alignments gave 243 ICE regions (median 22.6 kb; Table S9). Human ICE-positive genomes most often matched ICESsu-type references (79–80 of 117 genomes) and Tn916-type references (54 genomes), whereas the 12 ICE-positive fish genomes most often matched ICESde3396, a *Streptococcus dysgalactiae*-derived element, and the related ICESa2603-type references (9 genomes).

Resistance genes lay entirely within an ICE region in 92 of 159 human genomes (57.9%) but in only 4 of 100 fish genomes (4.0%; OR = 32.96, P = 4.4×10^{-21}; Table 5; Table S9). Of 325 resistance gene copies, 178 (55%) lay inside an ICE region, compared with 5% of random positions on the same contigs, and the proportion differed markedly among genes (Table 5): *aadE* (15/15 copies), *tet*(M) (77/95; 81%) and *tet*(O) (18/24; 75%) were mostly ICE-located; *aph*(3')-IIIa (69%), *erm*(B) (51%) and *mef*(A) and *msr*(D) (50% each) about half; whereas *erm*(A), *spw*, *lsa*(E) and *lnu*(B) (none of 43 copies) were never inside an ICE region. Regions carrying *tet*(M) matched the Tn916 reference in 57 of 77 cases (54 of 74 human, 3 of 3 fish), whereas none of the regions carrying *erm*(B), *tet*(O), *aadE* or *mef*(A) did. In fish, ICE-located resistance genes occurred in four genomes: *tet*(M) in 15.5–18.1-kb Tn916-like regions of an ST283 genome from Brazil, an ST1 genome from Bangladesh and an ST7 genome from Mexico, and *ant*(6)-Ia, *aph*(3')-IIIa and *tet*(S) in a 45-kb region of an ST7 genome from Vietnam. ICESde3396 was found in nine fish genomes (all ST7) and, in eight of them, no resistance gene lay in any ICE region. The macrolide efflux determinants *mef*(A) and *msr*(D) were detected only in human genomes.

Table 5. Acquired resistance genes located inside ICE regions. For each gene family: the genomes carrying it, the genomes in which at least one copy lies entirely inside an ICE region, and the share of all gene copies inside ICE regions. Gene families found in at most two genomes are combined. ICE regions are listed in Table S9.

Gene	Antibiotic class	Human genomes carrying	Human genomes, gene in ICE region	Fish genomes carrying	Fish genomes, gene in ICE region	Gene copies in ICE regions / total (%)
*aad*E	Aminoglycoside	15	15	0	0	15/15 (100%)
tet(M)	Tetracycline	91	74	3	3	77/95 (81%)
tet(O)	Tetracycline	24	18	0	0	18/24 (75%)
aph(3')-IIIa	Aminoglycoside	15	10	1	1	11/16 (69%)
erm(B)	Lincosamide/macrolide/streptogramin	40	22	2	0	22/43 (51%)
mef(A)	Macrolide	21	11	0	0	11/22 (50%)
msr(D)	Macrolide/streptogramin	20	10	0	0	10/20 (50%)
*cat*A16	Phenicol	5	2	0	0	2/5 (40%)
ant(6)-Ia	Aminoglycoside	22	10	1	1	11/28 (39%)
tet(S)	Tetracycline	2	0	1	1	1/3 (33%)
erm(A)	Lincosamide/macrolide	15	0	1	0	0/16 (0%)
spw	Aminoglycoside	10	0	0	0	0/10 (0%)
lsa(E)	Lincosamide/streptogramin	9	0	0	0	0/9 (0%)
lnu(B)	Lincosamide	8	0	0	0	0/8 (0%)
Other genes (8 families)	—	—	—	—	—	0/11 (0%)
Any resistance gene		122	92	6	4	178/325 (55%)

Discussion

Antimicrobial resistance in *S. agalactiae* has been studied largely within the human clinical compartment, leaving open the question of whether the same mobile genetic infrastructure that shaped human GBS evolution also operates in the aquaculture reservoirs with which humans increasingly share both geography and antimicrobial exposure. By analyzing 259 genomes from human and fish hosts across 25 countries within a single framework combining pangenome reconstruction, host-association analysis, and coordinate-resolution mobilome characterization, this study demonstrates that the two host compartments are separated not merely by lineage but by fundamentally distinct mobile element

ecologies a distinction with direct consequences for how resistance is carried, distributed, and potentially exchanged at the human–aquaculture interface. The findings develop as a connected argument: from a quantitative asymmetry in mobile element burden, through the mechanistic basis and clinical implications of that asymmetry, to the specific lineages and geographies where the two compartments converge.

The pronounced ICE prevalence asymmetry between human and fish-associated GBS is compatible with differential antimicrobial selective pressure, although in this dataset host is almost completely confounded with clonal background (fish genomes fall into two lineages), so selection and ancestry cannot be separated. Because ICEs encode their own excision, conjugation, and re-integration machinery and progressively accumulate cargo genes across lineage and species boundaries [43, 44], their disproportionate concentration in human GBS may reflect sustained selective enrichment rather than an incidental feature of clonal background. This interpretation is anchored in an established precedent: Da Cunha et al. showed that Tn916-borne tet(M) acquisition drove the mid-twentieth-century clonal replacement that fixed the modern human-adapted GBS population under tetracycline pressure [14]. Our results are consistent with this model: the ICE-rich architecture it produced is present in contemporary human GBS from five continents, whereas fish-associated lineages carry far fewer ICEs, and far fewer ICE-located resistance genes, despite substantial antimicrobial use in aquaculture [7]. The concordant prophage asymmetry (human 100% versus fish 54%; OR=∞; $P = 2.77\times10^{-23}$), consistent with the documented role of prophage content in GBS host adaptation [45, 46], and the host-partitioned insertion-sequence repertoires observed here, indicate that this divergence extends across the entire mobilome rather than being confined to ICEs. Collectively, these observations establish that human and fish GBS occupy divergent mobilome ecologies, and they raise the immediate question of whether this divergence has left the first-line therapeutic vulnerability of the species intact.

Among the most clinically significant findings of this study is the universal absence of β-lactam resistance genes across all 259 genomes, encompassing human clinical isolates from 15 countries and fish-associated isolates from 14 countries including regions of intensive aquaculture and substantial antibiotic use. This finding's significance lies precisely in its uniformity, because penicillin G remains the sole recommended agent for GBS intrapartum prophylaxis and neonatal therapy worldwide [47, 48, 49]. The result is consistent with the long-standing clinical observation that reduced penicillin susceptibility in GBS arises only sporadically through pbp2x point mutations and has never been mobilized or achieved epidemiological spread [50, 51], and it is corroborated by recent long-term surveillance reporting uniform penicillin susceptibility across invasive isolates [52]. Our data extend this reassurance to the aquaculture compartment, where no β-lactam determinant was detected despite intensive antimicrobial exposure. This finding only partially aligns with reports of phenotypic β-lactam resistance in aquaculture GBS from India [53]; the discrepancy is most plausibly explained not by contradiction but by methodological complementarity, since phenotypic resistance may arise from mechanisms chromosomal PBP alterations or regional selection that acquired-resistance gene databases do not capture, underscoring that genomic and phenotypic surveillance are non-redundant. The biological and public-health import is twofold: the therapeutic cornerstone of GBS management remains intact across both compartments, yet the demonstrated ICE-carrying capacity of human lineages means that mobilization of a β-lactam determinant onto an ICE backbone directly analogous to the historical tet(M) sweep [14] would represent a qualitatively new and urgent threat.

Coordinate-level analysis shows that ICE regions hold most of the acquired resistance genes in human GBS: 57.9% of human genomes carried at least one resistance gene inside an ICE region, and 55% of all resistance gene copies (81% of *tet*(M)) lay inside ICE regions, against 5% expected by chance. The association is gene-specific, however: *tet*(M), *tet*(O) and *aadE* were mostly ICE-located, *erm*(B) and the *mef*(A)/*msr*(D) pair about half, and *erm*(A), *spw*, *lsa*(E) and *lnu*(B) never, so these genes must reside on other elements that were not identified here; and location inside an ICE region is evidence of co-location, not by itself of mobilization. Tn916-like elements were the main carriers of *tet*(M) (57 of 77 *tet*(M)-carrying regions matched the Tn916 reference), consistent with early molecular characterizations in which *tet*(M) predominated among tetracycline-resistant GBS and *erm*(B) was the principal macrolide determinant [54], whereas

erm(B), *tet*(O), *aadE* and *mef*(A) sat in ICE regions of other types. *erm*(B) and *tet*(M) lay together inside a single ICE region in 16 human genomes; such physical linkage on one transferable element permits co-dissemination of macrolide and tetracycline resistance as a unit [55, 56, 57], a mechanism of composite ICE assembly now experimentally demonstrated in GBS through the transfer of a multidrug-resistance composite ICE from *Streptococcus anginosus* [58] and through site-specific nested integration of Tn1806 within ICESa2603-family elements [59]. ICE-located resistance was not confined to the human compartment: four fish genomes carried resistance genes inside ICE regions, three of them *tet*(M) on Tn916-like elements.

The fish mobilome is not a diminished copy of the human mobilome. ICESde3396, originally characterized in *Streptococcus dysgalactiae* subsp. *equisimilis* and shown to transfer across β-hemolytic streptococci including GBS [60, 61], was found in nine fish ST7 genomes, and in eight of them no resistance gene lay in any ICE region. ICE-located resistance in fish instead occurred in four genomes of three sequence types: Tn916-like regions carrying *tet*(M) in ST283 (Brazil), ST1 (Bangladesh) and ST7 (Mexico) genomes, and a 45-kb region carrying *ant*(6)-Ia, *aph*(3')-IIIa and *tet*(S), which also matches ICESde3396 references, in an ST7 genome from Vietnam. The low frequency of ICE-located resistance in fish (4.0% of genomes versus 57.9% in human) concords with regional phenotypic surveillance describing fish ST7 GBS as predominantly antibiotic-susceptible [62] and with the conserved core-genomic architecture of ST7 across Asian and Latin American aquaculture [63]. A plausible interpretation is that fish ST7 lineages have retained conjugative competence while remaining largely outside the sustained resistance-selective regime that shaped human GBS, since SDSE-to-GBS transfer of resistance-bearing ICEs has been demonstrated experimentally [64]. The presence of ICE-borne resistance in a small number of fish genomes shows that it already occurs in aquaculture lineages, while antimicrobial use in tilapia farming continues to expand across Asia and Africa [7].

Three genomic observations from this dataset identify specific lineages and geographies as priorities for targeted One Health surveillance, while remaining explicitly hypothesis-generating rather than conclusive given their basis in single isolates. First, a fish-associated genome from Bangladeshi aquaculture (GCF_031584935.1) was assigned ST1, a lineage belonging to CC1 that is widely distributed among human invasive GBS disease globally [65, 66], with serotype V predominance in adult infections [65]; it grouped with three human ST1 genomes from Houston, USA, in a four-genome clade supported at 100% bootstrap (77% for its sister relationship to two of them), and carried *tet*(M) inside a Tn916-like ICE region. Second, a Brazilian fish ST283 genome, belonging to the zoonotic lineage responsible for the 2015 Singapore outbreak [8, 9] and subsequently documented in human cases in Malaysia [67] and across Southeast Asia [68], also carried *tet*(M) inside a Tn916-like ICE region, although ST283 has been reported to undergo net loss rather than gain of resistance genes during host transitions [69]. Third, an ST7 genome from Mexico carried *tet*(M) in a Tn916-like region; the three fish elements are of similar size (15.5–18.1 kb) and match the same references as human Tn916-like elements, but this similarity does not establish transfer, because Tn916-like elements are widespread across streptococci. That these observations concentrate in South Asia and Latin America regions of intensive aquaculture, high human GBS burden, and minimal genomic surveillance, and where a novel ST7 variant is currently expanding [70] indicates that the human–aquaculture interface is localized to identifiable lineages and geographies where targeted, matched-sampling surveillance is most warranted.

Several limitations of this study warrant explicit acknowledgment. The geographic distribution of publicly available fish-associated GBS genomes reflects current global sequencing investment rather than disease burden: Brazil contributed 41% of the fish dataset, predominantly comprising ST260 and serotype Ib lineages, while systematic genomic surveillance of fish GBS from Africa, South Asia, and Southeast Asia where tilapia farming is most intensive and antimicrobial use least regulated remains essentially absent from public repositories. The present study was designed as a binary human–fish comparative analysis; bovine-associated GBS genomes were deliberately excluded as they represent a phylogenetically distinct host-adapted lineage with limited epidemiological overlap with the human–aquaculture interface investigated here. ICE detection relied on BLASTn alignment (≥ 15 kb, ≥ 90% identity) to ICEberg 2.0

references rather than on de novo structural annotation, so divergent or truncated elements and ICE types absent from the database are probably missed, and a resistance gene outside a detected ICE region is not necessarily outside a mobile element. The three single-genome interface observations ST1 Bangladesh, ST283 Brazil, and ST7 Mexico are hypothesis-generating rather than conclusive and require validation through targeted surveillance studies. The pangenome was defined using a strict core threshold of 99% genome prevalence, yielding a higher-confidence but smaller core genome than the more commonly used 95% threshold. Finally, host is confounded with lineage in this dataset (fish genomes fall into two lineages), so the host associations reported here describe differences between these populations and cannot be separated from ancestry.

Conclusion

In 259 *Streptococcus agalactiae* genomes, human and aquaculture populations differed sharply in their mobile element content: integrative conjugative elements were present in 73.6% of human but 12.0% of fish genomes, and resistance genes lay inside ICE regions in 57.9% of human but only 4.0% of fish genomes. ICE regions held most acquired resistance gene copies (55%; 81% of tet(M)), although erm(A), spw, lsa(E) and lnu(B) were never found within them, so ICEs are major but not exclusive carriers of resistance in this species. No β-lactam resistance gene was detected in either host, leaving first-line therapy unaffected in this dataset. In fish, ICE-located resistance was confined to four genomes, three carrying tet(M) on Tn916-like elements; these single-isolate observations are hypothesis-generating and do not establish transfer between hosts. Because host is closely tied to lineage in public genome collections, matched sampling of human and aquaculture isolates from the same regions, together with phenotypic susceptibility testing, is the logical next step for One Health surveillance of resistance at the human–aquaculture interface.

Declarations

Data availability

All genome assemblies analyzed in this study are publicly available from the NCBI Assembly database (accession numbers, BioSample identifiers, host, country of origin, sequence type and serotype metadata are provided in Supplementary Table S1). The processed data underlying all figures and tables are provided in Supplementary Tables S1–S9.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study is a secondary analysis of bacterial whole-genome sequences and associated metadata (host category, country of isolation, isolation source) that were publicly deposited in the NCBI Assembly and BioSample databases by the original investigators. The authors did not recruit participants, collect clinical samples or perform experiments on animals, and no personally identifiable information was accessed. Ethical approval and informed consent were therefore not required for this work. Ethical approval and consent for the original collection of isolates are the responsibility of the depositing studies. The accession numbers of all genomes analyzed are listed in Supplementary Table S1.

Author contributions

Sayed Mashequl Bari: Conceptualization, methodology, data curation, formal analysis, visualization, writing - original draft. **Fouzia Mostafa:** Data curation, formal analysis, writing - original draft. **Sabrina Hossain**: Writing - review & editing. **Md. Abdullah-Al-Mamun**: Writing - review & editing. **Muhammad Manjurul Karim**: Writing - review & editing and supervision. All authors reviewed and approved the final manuscript.

References

1. Phares CR, Lynfield R, Farley MM, et al. Epidemiology of invasive group B streptococcal disease in the United States, 1999–2005. JAMA. 2008;299:2056–65. https://doi.org/10.1001/jama.299.17.2056.
2. Lawn JE, Bianchi-Jassir F, Russell NJ, et al. Group B streptococcal disease worldwide for pregnant women, stillbirths, and children: why, what, and how to undertake estimates? Clin Infect Dis. 2017;65(Suppl 2):S89–99. https://doi.org/10.1093/cid/cix653.
3. Russell NJ, Seale AC, O'Sullivan C, et al. Risk of early-onset neonatal group B streptococcal disease with maternal colonization worldwide: systematic review and meta-analyses. Clin Infect Dis. 2017;65(Suppl 2):S152–9. https://doi.org/10.1093/cid/cix655.
4. Nuccitelli A, Rinaudo CD, Maione D. Group B Streptococcus vaccine: state of the art. Ther Adv Vaccines. 2015;3:76–90. https://doi.org/10.1177/2051013615579869.
5. Sheehan KB, Martin M, DeSousa C, et al. Identification and characterization of a second tilapia-adapted *S. agalactiae* ST261 outbreak causing mass mortality in aquaculture farms. Front Microbiol. 2009;10:1–12. https://doi.org/10.3389/fmicb.2009.00001.
6. Liu G, Zhang W, Lu C. Comparative genomics analysis of *Streptococcus agalactiae* reveals that pathogenic and commensal strains differ in their cell surface components. BMC Genomics. 2013;14:1. https://doi.org/10.1186/1471-2164-14-1.
7. Schar D, Klein EY, Laxminarayan R, Gilbert M, Van Boeckel TP. Global trends in antimicrobial use in aquaculture. Sci Rep. 2020;10:21878. https://doi.org/10.1038/s41598-020-78849-3.
8. Kalimuddin S, Chen SL, Lim CTK, et al. 2015 epidemic of severe *Streptococcus agalactiae* sequence type 283 infections in Singapore associated with the consumption of raw freshwater fish. Clin Infect Dis. 2017;64(Suppl 2):S145–52. https://doi.org/10.1093/cid/cix021.
9. Barkham T, Zadoks RN, Azmai MNA, et al. One hypervirulent clone, sequence type 283, accounts for a large proportion of invasive *Streptococcus agalactiae* isolated from humans and diseased tilapia in Southeast Asia. PLOS Negl Trop Dis. 2019;13:e0007421. https://doi.org/10.1371/journal.pntd.0007421.
10. Crestani C, Forde TL, Lycett SJ, et al. The fall and rise of group B Streptococcus in dairy cattle: reintroduction due to human-to-cattle host jumps? Microb Genomics. 2021;7:000648. https://doi.org/10.1099/mgen.0.000648.
11. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. Proc Natl Acad Sci USA. 2005;102:13950–5. https://doi.org/10.1073/pnas.0506758102.
12. Brochet M, Couvé E, Zouine M, et al. Genomic diversity and evolution within the species *Streptococcus agalactiae*. Microbes Infect. 2006;8:1227–43. https://doi.org/10.1016/j.micinf.2005.11.010.
13. Brochet M, Couvé E, Glaser P, Guédon G, Payot S. Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. J Bacteriol. 2008;190:6913–7. https://doi.org/10.1128/jb.00824-08.
14. Da Cunha V, Davies MR, Douarre PE, et al. *Streptococcus agalactiae* clones infecting humans were selected and fixed through the extensive use of tetracycline. Nat Commun. 2014;5:4544. https://doi.org/10.1038/ncomms5544.

15. Gori A, Harrison OB, Mlia E, et al. Pan-GWAS of *Streptococcus agalactiae* highlights lineage-specific genes associated with virulence and niche adaptation. mBio. 2020;11:e00728–20. https://doi.org/10.1128/mBio.00728-20.
16. Crestani C, Forde TL, Bell J, et al. Genomic and functional determinants of host spectrum in group B *Streptococcus*. PLOS Pathog. 2024;20:e1012400. https://doi.org/10.1371/journal.ppat.1012400.
17. Rosinski-Chupin I, Sauvage E, Mairey B, et al. Reductive evolution in *Streptococcus agalactiae* and the emergence of a host adapted lineage. BMC Genomics. 2013;14:252. https://doi.org/10.1186/1471-2164-14-252.
18. Richards VP, Lang P, Pavinski Bitar PD, et al. Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. Infect Genet Evol. 2011;11:1311–22. https://doi.org/10.1016/j.meegid.2011.04.019.
19. NCBI, Datasets. CLI v18.36.0. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/datasets/
20. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3. https://doi.org/10.1093/bioinformatics/btp163.
21. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022;38:5315–6. https://doi.org/10.1093/bioinformatics/btac672.
22. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods. 2023;20:1203–12. https://doi.org/10.1038/s41592-023-01940-w.
23. Orakov A, Fullam A, Coelho LP, et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol. 2021;22:178. https://doi.org/10.1186/s13059-021-02393-0.
24. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962. https://doi.org/10.1371/journal.pone.0163962.
25. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8. https://doi.org/10.1038/ismej.2017.126.
26. Schwengers O, Jelonek L, Dieckmann MA, et al. Bakta: rapid and standardized annotation of bacterial genomes via a comprehensive database. Microb Genomics. 2021;7:000685. https://doi.org/10.1099/mgen.0.000685.
27. Seemann T mlst, editor. GitHub. https://github.com/tseemann/mlst
28. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.
29. Chen SL. GBS-SBG: group B *Streptococcus* serotyping by genome. GitHub. https://github.com/swainechen/GBS-SBG
30. Tonkin-Hill G, MacAlasdair N, Ruis C, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21:180. https://doi.org/10.1186/s13059-020-02090-4.
31. Gu Z. Complex heatmap visualization. iMeta. 2022;1:e43. https://doi.org/10.1002/imt2.43.
32. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15. https://doi.org/10.1093/nar/gku1196.
33. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and methods for phylogenetic inference. Mol Biol Evol. 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015.
34. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114. https://doi.org/10.1038/s41467-018-07641-9.

35. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36. https://doi.org/10.1111/2041-210X.12628.
36. Feldgarden M, Brover V, Gonzalez-Escalona N, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11:12728. https://doi.org/10.1038/s41598-021-91456-0.
37. Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. 2022;50:D912–7. https://doi.org/10.1093/nar/gkab1107.
38. Lees JA, Vehkala M, Välimäki N et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*. 2016;7:12797. https://doi.org/10.1038/ncomms12797; Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. Pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*. 2018;34:4310–4312. https://doi.org/10.1093/bioinformatics/bty539.
39. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics. 2017;33:3340–7. https://doi.org/10.1093/bioinformatics/btx433.
40. Liu M, Li X, Xie Y, et al. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic Acids Res. 2019;47:D660–5. https://doi.org/10.1093/nar/gky1123.
41. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res. 2012;40:e126. https://doi.org/10.1093/nar/gks406.
42. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genomics. 2018;4:e000206. https://doi.org/10.1099/mgen.0.000206.
43. Johnson CM, Grossman AD. Integrative and conjugative elements (ICEs): what they do and how they work. Annu Rev Genet. 2015;49:577–601. https://doi.org/10.1146/annurev-genet-112414-055018.
44. Ambroset C, Coluzzi C, Guédon G, et al. New insights into the classification and integration specificity of *Streptococcus* integrative conjugative elements through extensive genome exploration. Front Microbiol. 2015;6:1483. https://doi.org/10.3389/fmicb.2015.01483.
45. Crestani C, Forde TL, Zadoks RN. Development and application of a prophage integrase typing scheme for group B *Streptococcus*. Front Microbiol. 2020;11:1993. https://doi.org/10.3389/fmicb.2020.01993.
46. Wang R, Li L, Huang Y, et al. Pathogenicity of human ST23 *Streptococcus agalactiae* to fish and genomic comparison of pathogenic and non-pathogenic isolates. Front Microbiol. 2017;8:1933. https://doi.org/10.3389/fmicb.2017.01933.
47. Verani JR, McGee L, Schrag SJ. Prevention of perinatal group B streptococcal disease: revised guidelines from CDC, 2010. MMWR Recomm Rep. 2010;59(RR–10):1–36.
48. Prevention of group B streptococcal. early-onset disease in newborns: ACOG Committee Opinion, Number 797. Obstet Gynecol. 2020;135(2):e51–72. https://doi.org/10.1097/AOG.0000000000003668.
49. Puopolo KM, Lynfield R, Cummings JJ. Management of infants at risk for group B streptococcal disease. Pediatrics. 2019;144(2):e20191881. https://doi.org/10.1542/peds.2019-1881.
50. Kimura K, Wachino J, Kurokawa H, Suzuki S, Yamane K, Shibata N, Arakawa Y. Practical disk diffusion test for detecting group B streptococcus with reduced penicillin susceptibility. J Clin Microbiol. 2009;47:1107–9. https://doi.org/10.1128/jcm.02063-08.
51. Dahesh S, Hensler ME, Van Sorge NM, Gertz RE Jr, Schrag S, Nizet V, Beall BW. Point mutation in the group B streptococcal *pbp2x* gene conferring decreased susceptibility to β-lactam antibiotics. Antimicrob Agents Chemother. 2008;52:2915–8. https://doi.org/10.1128/aac.00461-08.

52. Joensen Ó, Krogfelt KA, Gaini S, Slotved HC. A nationwide study of invasive *Streptococcus agalactiae* in the Faroe Islands from 2009 to 2024. Sci Rep. 2026;16:5090. https://doi.org/10.1038/s41598-026-35025-3.
53. Kalathil Maniyappan K, Kalasseril Girijan S, Krishnan R, et al. Assessing multi-drug resistance in *Streptococcus agalactiae* infecting farmed Nile tilapia: findings from Kerala, India. Microb Pathog. 2025;205:107666. https://doi.org/10.1016/j.micpath.2025.107666.
54. Poyart C, Jardy L, Quesne G, Berche P, Trieu-Cuot P. Genetic basis of antibiotic resistance in *Streptococcus agalactiae* strains isolated in a French hospital. Antimicrob Agents Chemother. 2003;47:794–7. https://doi.org/10.1128/aac.47.2.794-797.2003.
55. Roberts AP, Mullany P. Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. FEMS Microbiol Rev. 2011;35:856–71. https://doi.org/10.1111/j.1574-6976.2011.00283.x.
56. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. Clin Microbiol Rev. 2018;31:e00088–17. https://doi.org/10.1128/cmr.00088-17.
57. Brenciani A, Bacciaglia A, Vecchi M, Vitali LA, Varaldo PE, Giovanetti E. Genetic elements carrying erm(B) in *Streptococcus pyogenes* and association with tet(M) tetracycline resistance gene. Antimicrob Agents Chemother. 2007;51:1209–16. https://doi.org/10.1128/aac.01484-06.
58. Wang Y, Liu T, Yi S, Zhu Y. A new composite integrative and conjugative element mediates multiple drug resistance from *Streptococcus anginosus* to *Streptococcus agalactiae*. Infect Drug Resist. 2025;18:1941–9. https://doi.org/10.2147/IDR.S514281.
59. Yi S, Xu X, Yin L, He Z, Wang X. Site-specific nested integration of Tn1806 into ICESa2603-family integrative and conjugative elements in *Streptococcus agalactiae*. Microorganisms. 2026;14:375. https://doi.org/10.3390/microorganisms14020375.
60. Davies MR, Shera J, Van Domselaar GH, Sriprakash KS, McMillan DJ. A novel integrative conjugative element mediates genetic transfer from group G *Streptococcus* to other β-hemolytic streptococci. J Bacteriol. 2009;191:4118–27. https://doi.org/10.1128/jb.01624-08.
61. Smyth DJ, Shera J, Bauer MJ, Cameron A, McNeilly CL, Sriprakash KS, McMillan DJ. Conjugative transfer of ICESde3396 between three β-hemolytic streptococcal species. BMC Res Notes. 2014;7:521. https://doi.org/10.1186/1756-0500-7-521.
62. Sapugahawatte DN, Li C, Dharmaratne P, Zhu C, Yeoh YK, Yang J, Lo NWS, Wong KT, Ip M. Prevalence and characteristics of *Streptococcus agalactiae* from freshwater fish and pork in Hong Kong wet markets. Antibiotics. 2022;11:397. https://doi.org/10.3390/antibiotics11030397.
63. Kayansamruaj P, Pirarat N, Kondo H, Hirono I, Rodkhum C. Genomic comparison between pathogenic *Streptococcus agalactiae* isolated from Nile tilapia in Thailand and fish-derived ST7 strains. Infect Genet Evol. 2015;36:338–44. https://doi.org/10.1016/j.meegid.2015.10.009.
64. Rochegüe T, Hughes S, Saras E, Gillet B, Drapeau A, Glaser P, Madec JY, Haenni M, Lupo A. Integrative and conjugative elements carrying high-level gentamicin resistance genes in *Streptococcus dysgalactiae* subsp. equisimilis from horses. Microb Genomics. 2026;12:001722. https://doi.org/10.1099/mgen.0.001722.
65. Teatero S, McGeer A, Low DE, Li A, Demczuk W, Martin I, Fittipaldi N. Characterization of invasive group B *Streptococcus* strains from the greater Toronto area, Canada. J Clin Microbiol. 2014;52:1441–7. https://doi.org/10.1128/JCM.03554-13.
66. Luan SL, Granlund M, Sellin M, Lagergård T, Spratt BG, Norgren M. Multilocus sequence typing of Swedish invasive group B streptococcus isolates indicates a neonatally associated genetic lineage and capsule switching. J Clin Microbiol. 2005;43:3727–33. https://doi.org/10.1128/JCM.43.8.3727-3733.2005.
67. Zohari Z, Barkham T, Maswan NM, Chen SL, Muthanna A, Lee KW, Mohd Desa MN, Amal Azmai MN, Joseph NMS, Amin-Nordin S. Fish-associated *Streptococcus agalactiae* ST283: first human cases reported from Malaysia. J Med

Microbiol. 2023;72:001729. https://doi.org/10.1099/jmm.0.001729.

68. Sirimanapong W, Phuoc NN, Crestani C, Chen S, Zadoks RN. Geographical, temporal and host-species distribution of potentially human-pathogenic group B Streptococcus in aquaculture species in Southeast Asia. Pathogens. 2023;12:525. https://doi.org/10.3390/pathogens12040525.
69. Schar D, Zhang Z, Pires J, Vrancken B, Suchard MA, Lemey P, Ip M, Gilbert M, Van Boeckel T, Dellicour S. Dispersal history and bidirectional human-fish host switching of invasive, hypervirulent *Streptococcus agalactiae* sequence type 283. PLOS Glob Public Health. 2023;3:e0002454. https://doi.org/10.1371/journal.pgph.0002454.
70. LaFrentz BR, Barato P, Keleher WR, et al. Emergence, identification, and characterization of a novel *Streptococcus agalactiae* CPS type Ia ST7 strain (Ia-2021) causing large scale mortalities in tilapia aquaculture. Sci Rep. 2026;16:20338. https://doi.org/10.1038/s41598-026-20338-x.

Figures

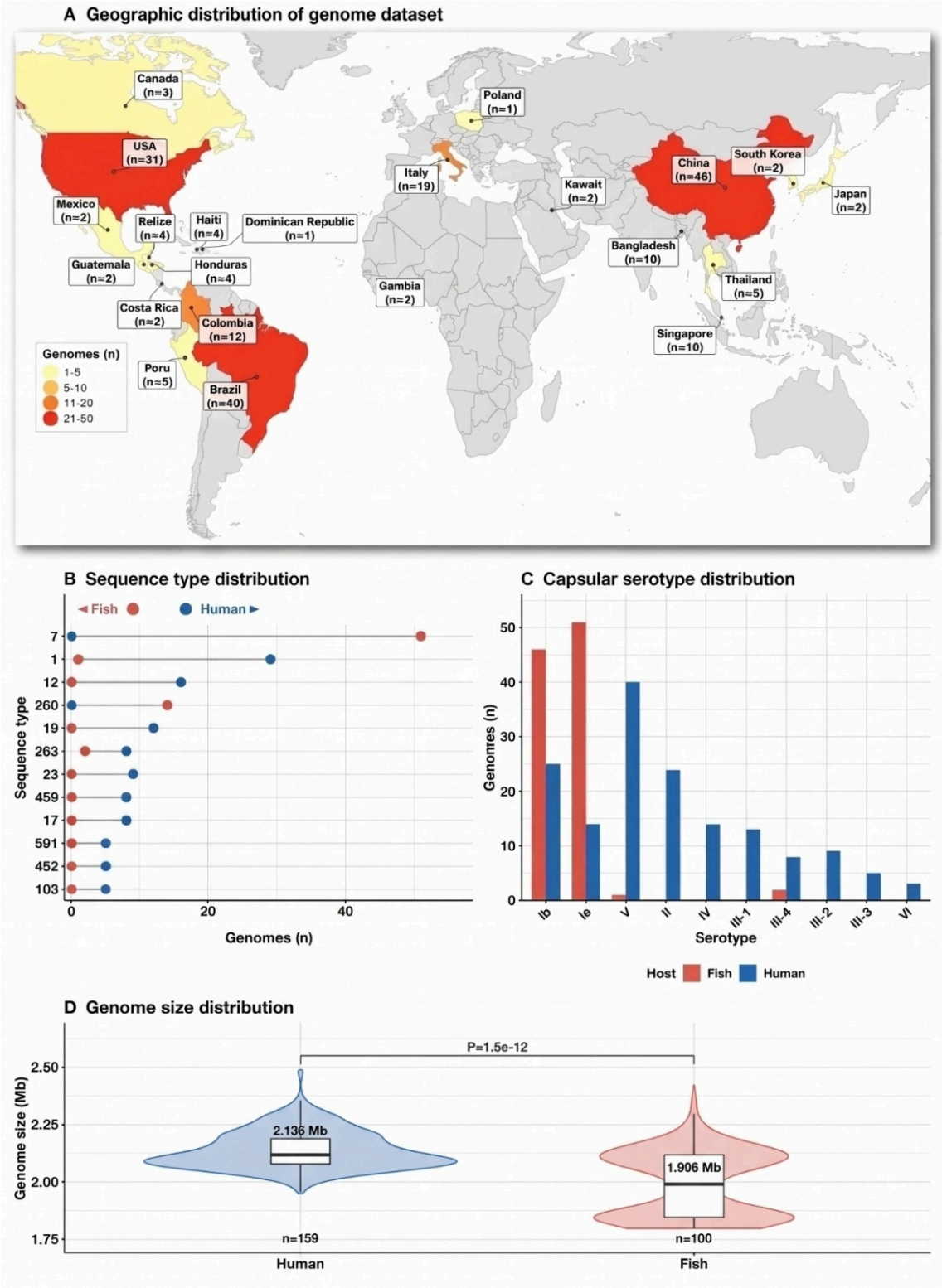


Figure 1

Overview of the study dataset. (A) Geographic distribution of 215 genomes with available country metadata (44 genomes lacked geographic data), shaded by genome count and colored by dominant host (red, fish; blue, human); labels shown for countries contributing ≥5 genomes. **(B)**Sequence type distribution by host. **(C)** Capsular serotype distribution by host. **(D)** Genome size distribution by host (Wilcoxon rank-sum test, $P=1.5\times10^{-12}$).

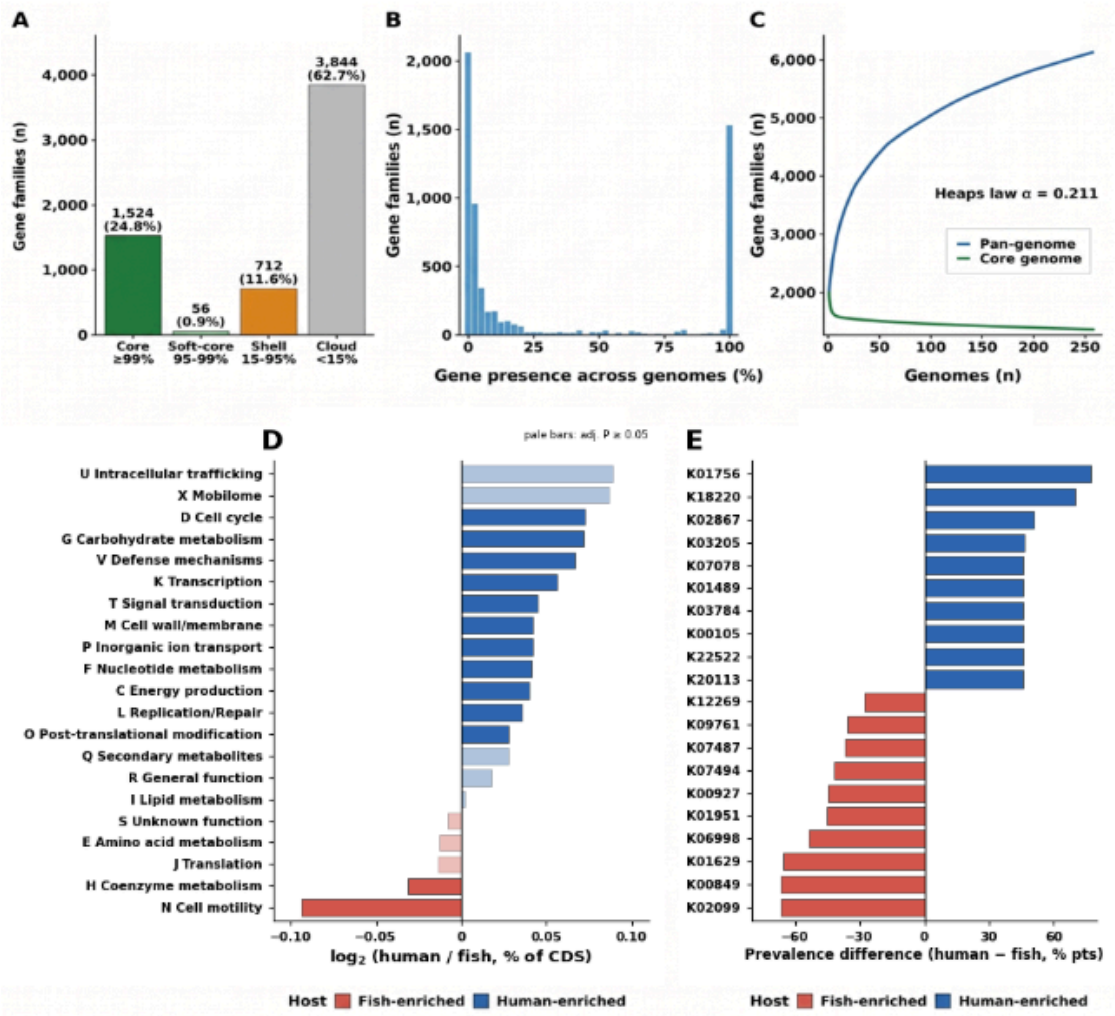


Figure 2

Pangenome architecture and host-differential functional content of *Streptococcus agalactiae*. (A) Partitioning of 6,136 gene families into core, soft-core, shell and cloud. **(B)** Gene-frequency distribution. **(C)** Pan-genome and core-genome accumulation curves; Heaps' law α = 0.211. **(D)** Difference in COG category composition ($\log_2$ ratio of the share of coding sequences, human/fish) for the 21 informative categories; filled bars, Benjamini–Hochberg adj. $P < 0.05$. **(E)** Host-differential KEGG Orthology terms by genome prevalence difference. n = 259 genomes throughout.

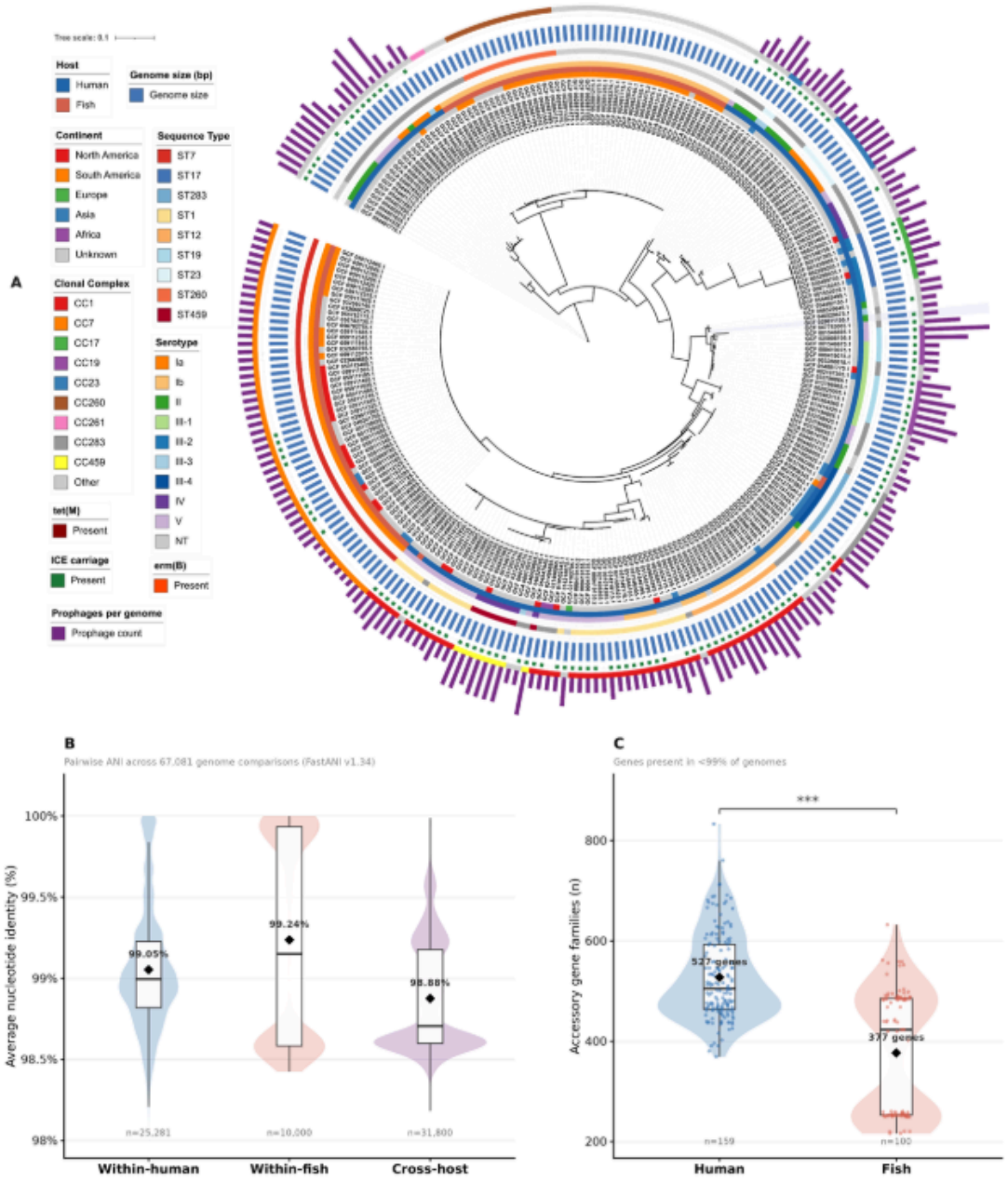


Figure 3

Core-genome phylogeny, genomic identity and accessory genome divergence of *Streptococcus agalactiae*. (A) Maximum-likelihood phylogeny; rings show continent, host, serotype, erm(B), sequence type, tet(M), genome size, ICE carriage, clonal complex, prophage count. **(B)** Pairwise ANI stratified by comparison type. **(C)** Accessory gene families per genome by host. n = 259 genomes.

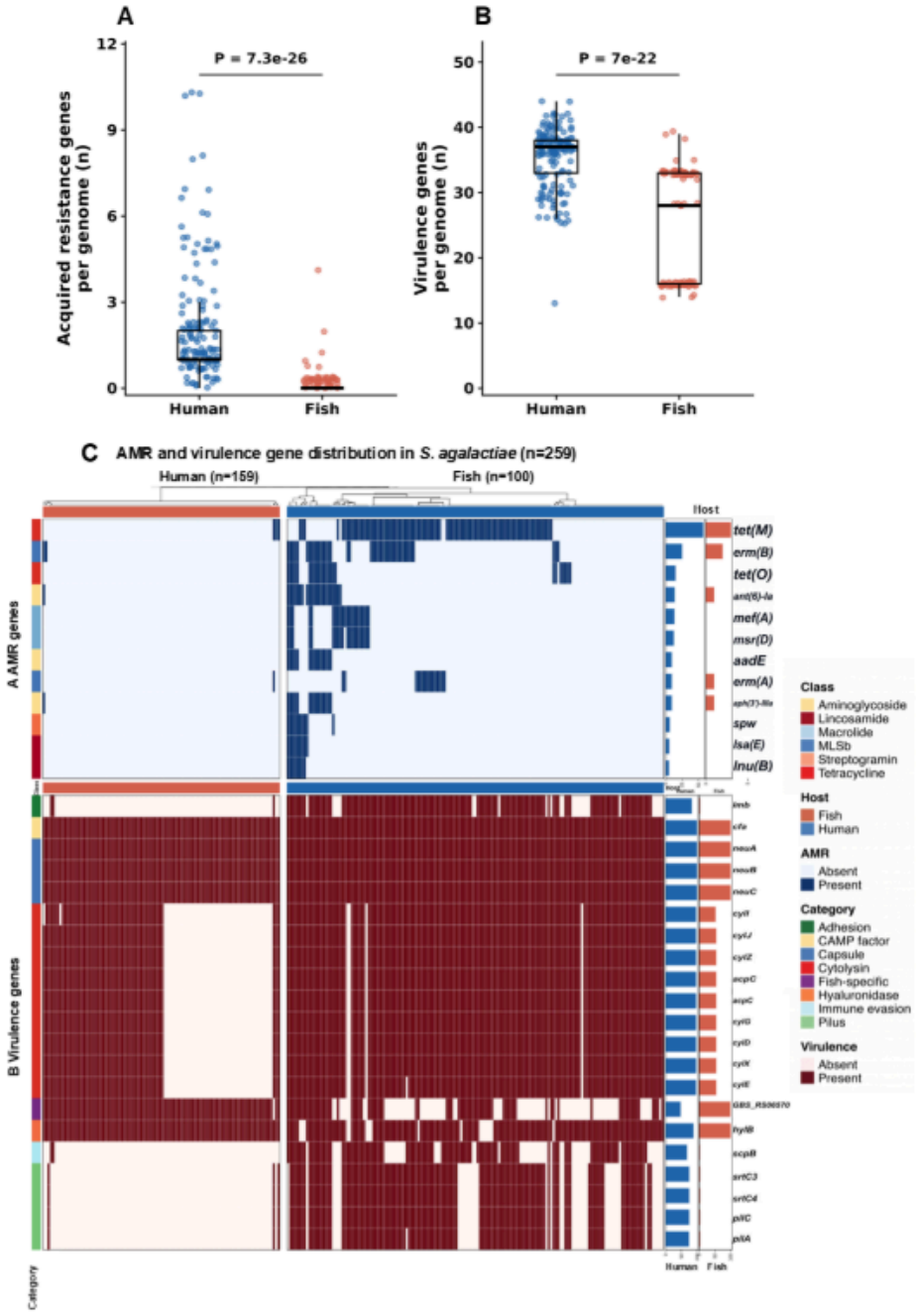


Figure 4

Resistance and virulence gene burden and distribution in *Streptococcus agalactiae*. (A) Acquired resistance genes per genome by host (P = 7.3 × 10^{-26}). **(B)**Virulence genes per genome by host (P = 7.0 × 10^{-22}). **(C)**Presence/absence of 12 significant resistance genes and 21 virulence genes across all genomes, clustered within host; bar charts show per-host prevalence. n = 259 genomes.

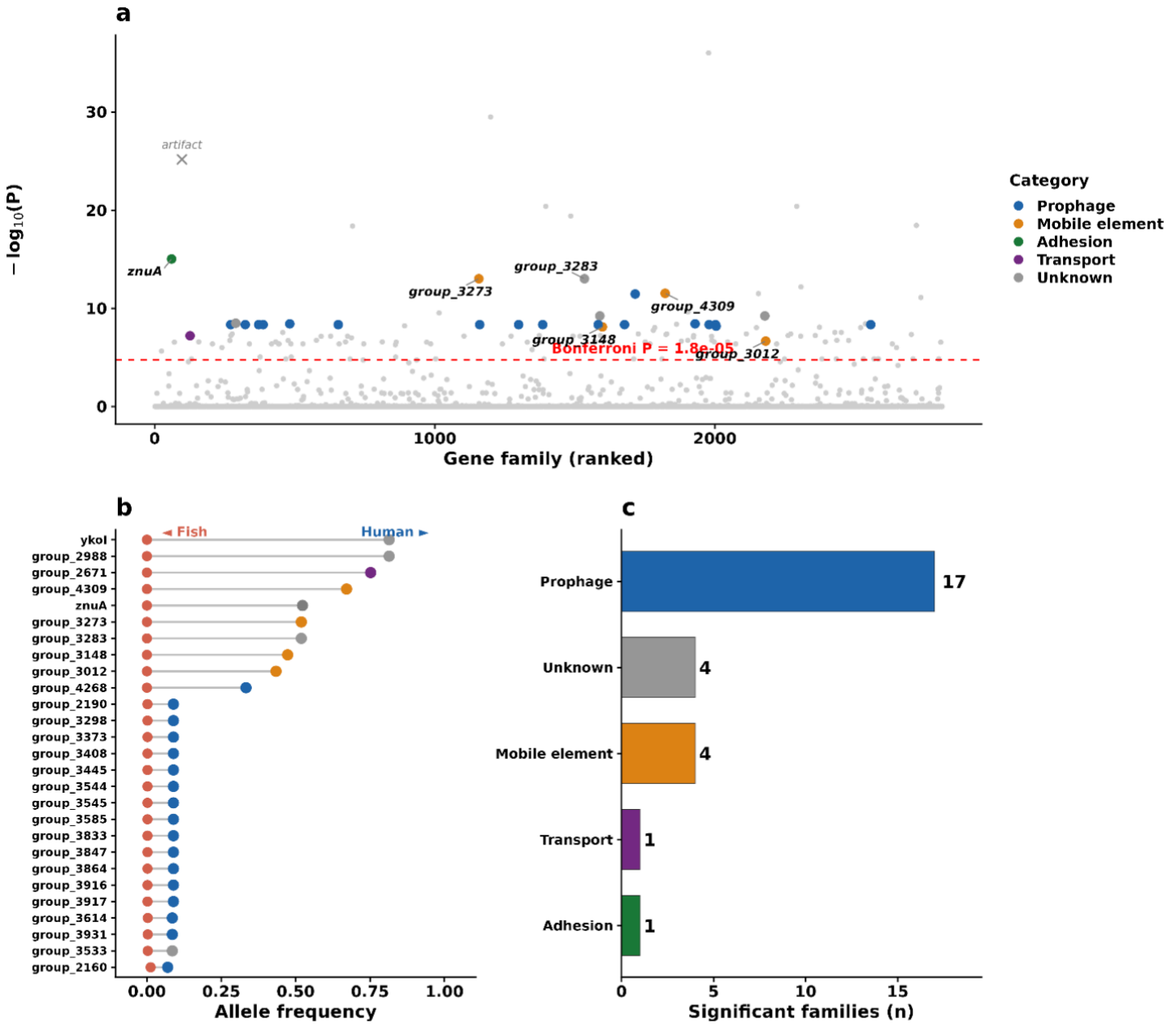


Figure 5

Pangenome-wide association analysis of host-associated gene families in *Streptococcus agalactiae*. (A) Manhattan plot of $-\log_{10}(P)$ for 2,811 gene families tested by linear mixed model (Pyseer v1.4.0); Bonferroni threshold ($P < 1.78\times10^{-5}$) shown as dashed red line; significant associations (n=27) coloured by functional category; artifact (group_2481) marked separately; lmb indicated. **(B)** Allele frequency of each significant gene family in human versus fish isolates; group_2481 excluded (see Table 4 note). **(C)**Number of significant gene families by functional category. All associations are human-enriched.

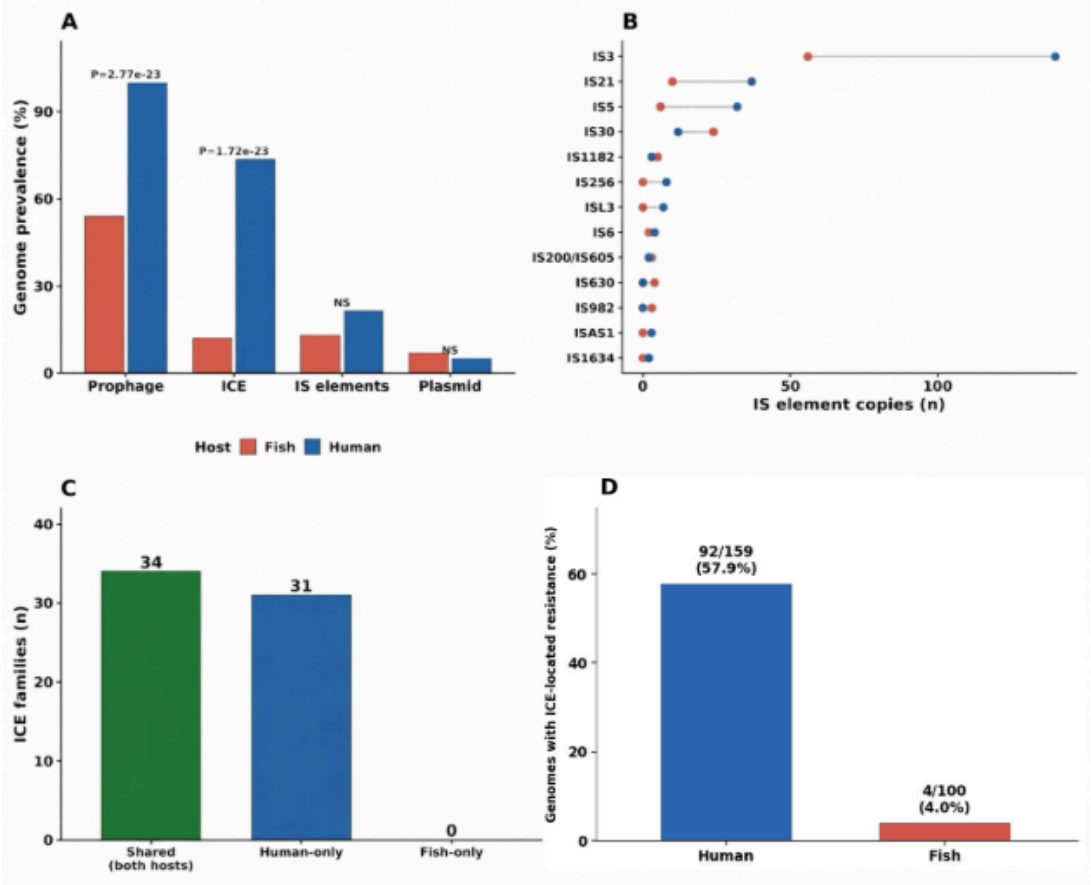


Figure 6

Mobilome architecture and resistance-element co-localization in *Streptococcus agalactiae*. (A) Genome prevalence of four mobile-element categories by host. **(B)** Insertion-sequence family distribution by host. **(C)** ICE families (ICEberg reference entries) detected in both hosts (34), in human genomes only (31) and in fish genomes only (none). **(D)** Genomes with at least one resistance gene inside an ICE region (human 92/159, fish 4/100). n = 259 genomes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesS1S9.xlsx